Discovery of Changes from the Global Carbon Cycle and Climate System Using Data Mining

Principal Investigator: Vipin Kumar, University of Minnesota

Co-Investigators:

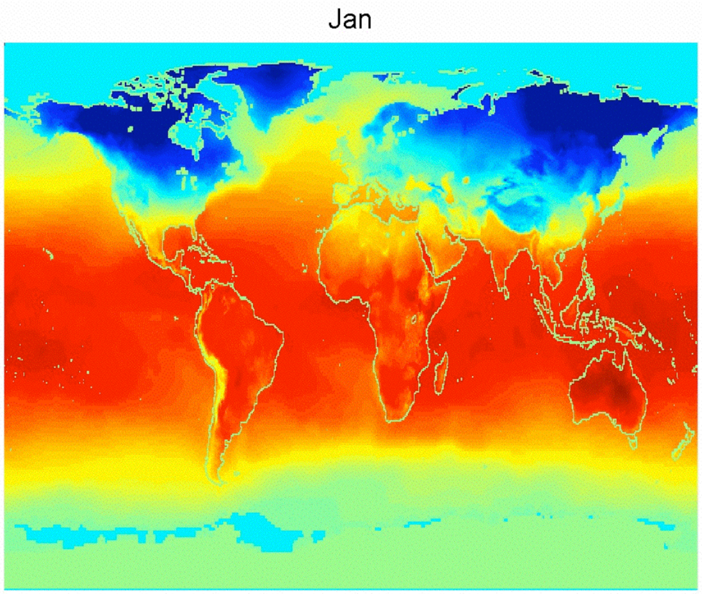
George Karypis	University of Minnesota
Steven Klooster	NASA Ames Research Center, UCMB
Shashi Shekhar	University of Minnesota
Christopher Potter	NASA Ames Research Center

Michael Steinbach, Pusheng Zhang, Varun Chandola, Gyorgy Simon, Shyam Boriah
University of Minnesota

Pang-Ning Tan
Michigan State University

Alicia Torregrosa, Vanessa Genovese
University of California, Monterrey Bay

Discovery of Patterns in the Earth Science Data

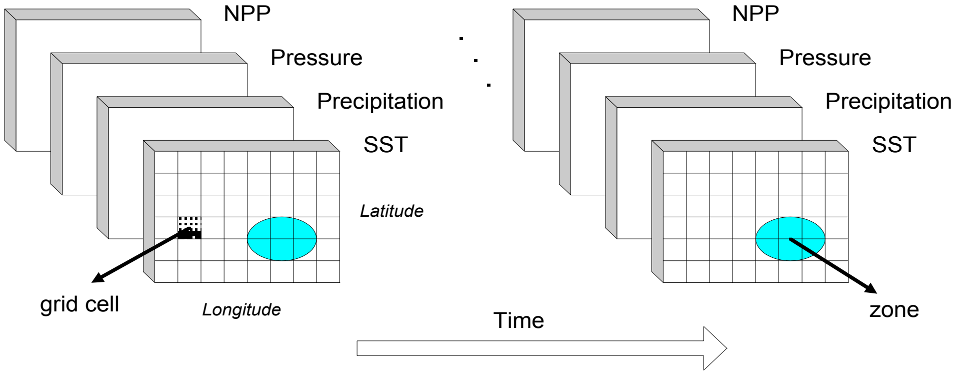


Land and Sea Temperature

Goal: Better understand global scale patterns in biosphere processes, especially relationships between the global carbon cycle and the climate system.

Challenge: Develop data mining techniques to efficiently find spatio-temporal patterns in large Earth Science data sets.

- Global snapshots of values for a number of variables on land surfaces or water
- Data sources:
  - weather observation stations
  - earth orbiting satellites (since 1981)
  - model-based data



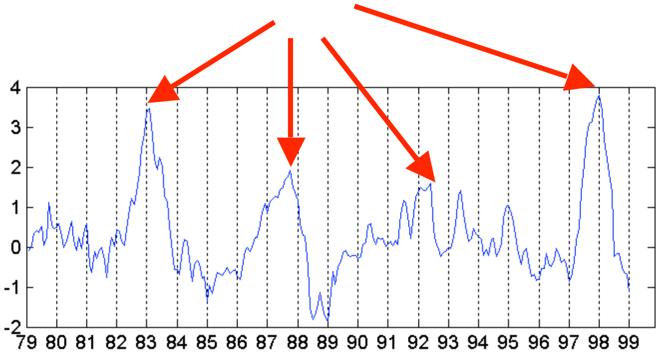
Contributions

- Clustering for the detection of climate indices
- Automated detection of ecosystem disturbances
- Association analysis to discover relationships between climate variables
- Efficient query processing for spatial time series

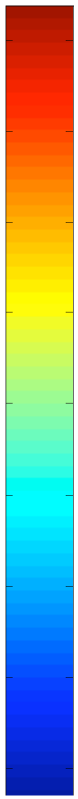
Climate Indices:
Connecting the Ocean/Atmosphere and the Land

- A climate index is a time series involving sea surface temperature or sea level pressure
- Climate indices can be used to explore teleconnections
  - The simultaneous variation in climate and related processes over widely separated points on the Earth

El Nino Events

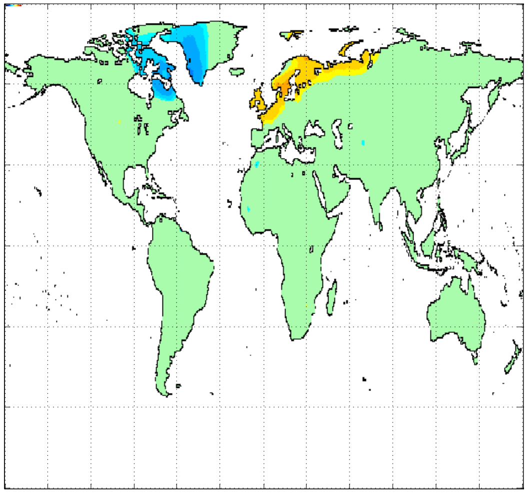


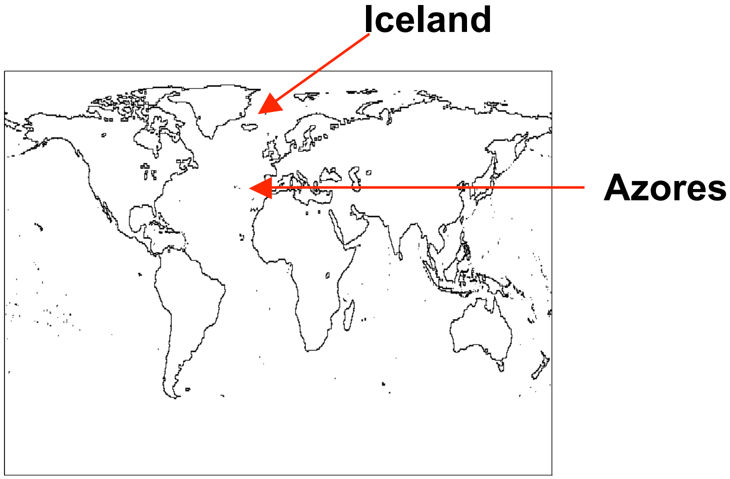
Anom 1+2 Index



Climate Indices - NAO

- The North Atlantic Oscillation (NAO) is associated with climate variation in Europe and North America.





- Normalized pressure differences between Ponta Delgada, Azores and Stykkisholmur, Iceland.
- Associated with warm and wet winters in Europe and in cold and dry winters in northern Canada and Greenland
- The eastern US experiences mild and wet winter conditions.

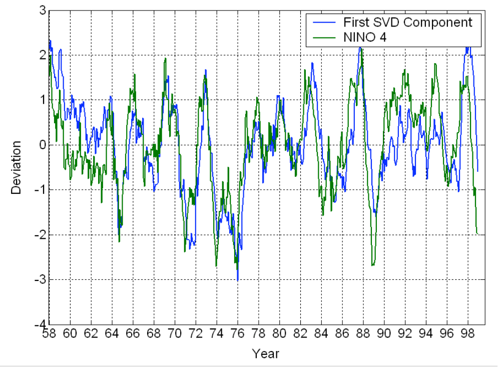
List of Well Known Climate Indices

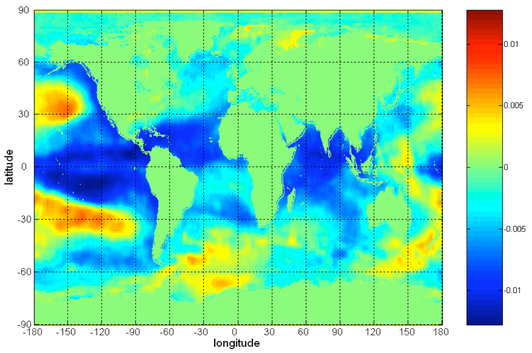
Index	Description
SOI	**Southern Oscillation Index:** Measures the SLP anomalies between Darwin and Tahiti
NAO	**North Atlantic Oscillation:** Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland
AO	**Arctic Oscillation:** Defined as the _first principal component of SLP poleward of 20° N
PDO	**Pacific Decadel Oscillation:** Derived as the leading principal component of monthly SST anomalies in the North Pacific Ocean, poleward of 20° N
QBO	**Quasi-Biennial Oscillation Index:** Measures the regular variation of zonal (i.e. east-west) strato-spheric winds above the equator
CTI	**Cold Tongue Index:** Captures SST variations in the cold tongue region of the equatorial Pacific Ocean (6° N-6° S, 180° -90° W)
WP	**Western Pacific:** Represents a low-frequency temporal function of the ‘zonal dipole' SLP spatial pattern involving the Kamchatka Peninsula, southeastern Asia and far western tropical and subtropical North Pacific
NINO1+2	Sea surface temperature anomalies in the region bounded by 80° W-90° W and 0° -10° S
NINO3	Sea surface temperature anomalies in the region bounded by 90° W-150° W and 5° S-5° N
NINO3.4	Sea surface temperature anomalies in the region bounded by 120° W-170° W and 5° S-5° N
NINO4	Sea surface temperature anomalies in the region bounded by 150° W-160° W and 5° S-5° N

Discovering Climate Indices

- Observation
  - The El Nino phenomenon was first noticed by Peruvian fishermen centuries ago as a relationship between a persistent warm southward current around Christmas and a disastrous impact on fishing.
- Eigenvalue techniques such as Principal Components Analysis (PCA/EOF) and Singular Value Decomposition (SVD).
  - Components (patterns) must be orthogonal making physical interpretation difficult.
  - Stronger patterns tend to hide weaker patterns
  - Requires domain knowledge to select the regions of interest

We applied SVD to the global Sea Surface Temperature (SST)

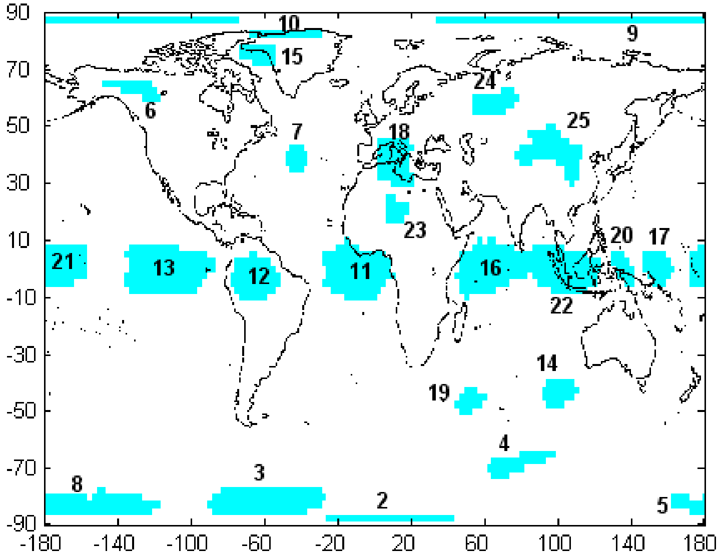




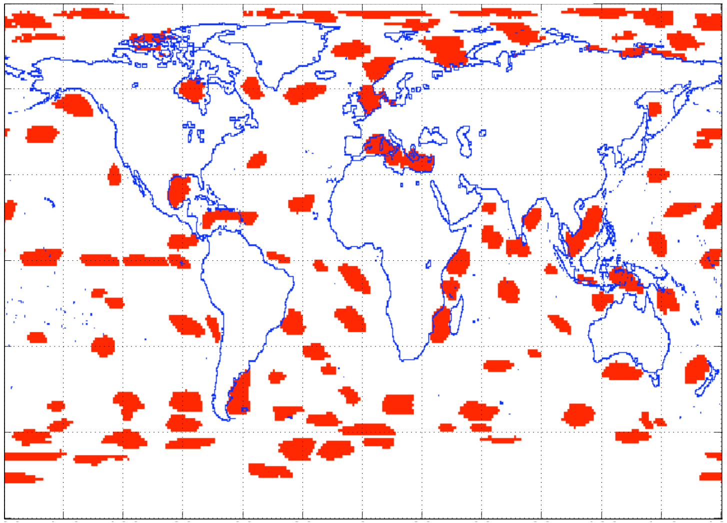
Discovering Climate Indices via Data Mining

- Clustering provides an alternative approach for finding candidate indices.
  - Clusters represent regions with relatively homogeneous behavior with respect to SST or SLP.
  - The centroids of these clusters are time series that summarize the behavior of these areas, and thus, represent potential climate indices.
- Shared Nearest Neighbor (SNN) clustering finds groups of points (SST or SLP time series, in this case) that have relatively homogeneous behavior.
  - Alleviates problems with varying density and problems with clusters of different shapes and sizes.
  - Can handle noisy data such as Earth Science data
  - Finds the number of clusters automatically

SLP and SST Clusters



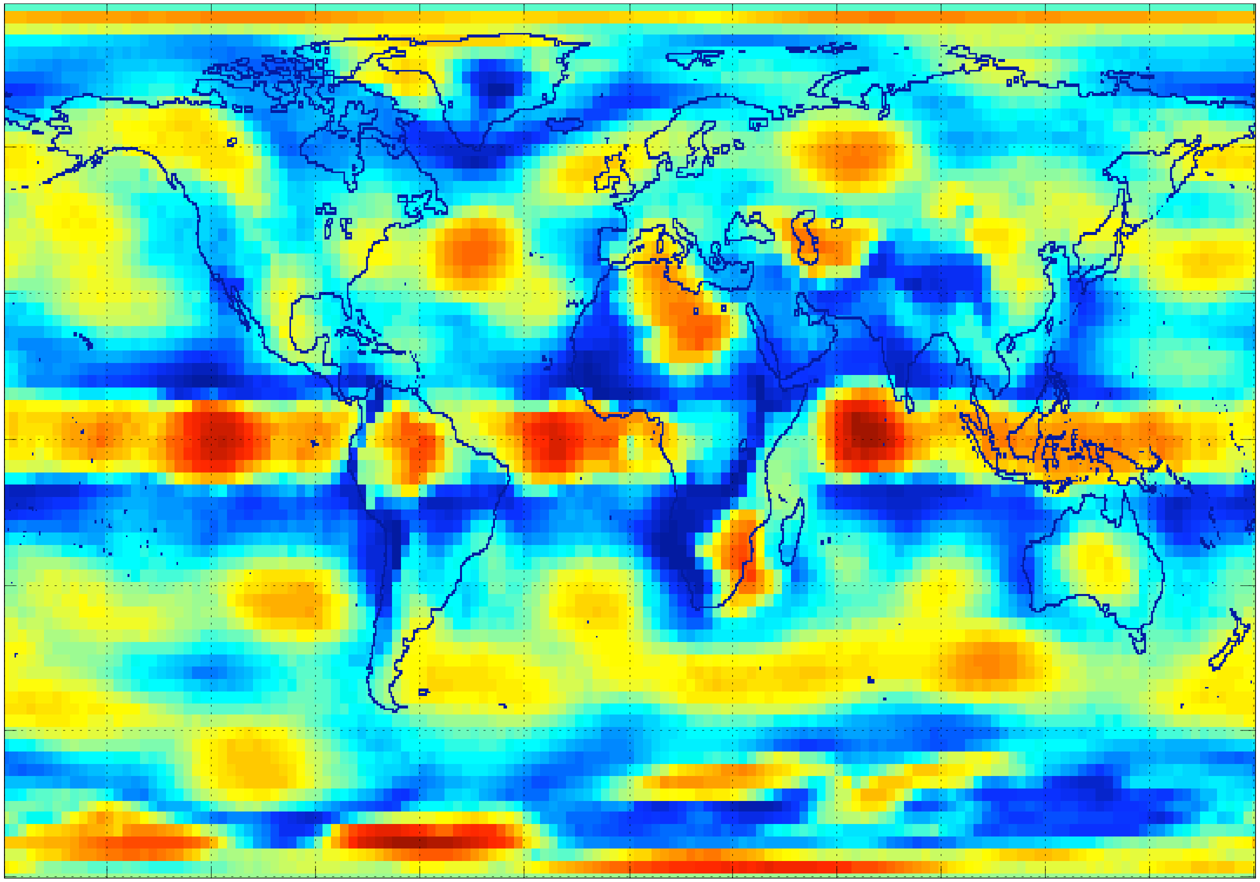
25 SLP Clusters



107 SST Clusters

Homogeneity of SLP Time Series

Our SNN clustering approach defines a density that measures the homogeneity of each time series with respect to its neighbors.

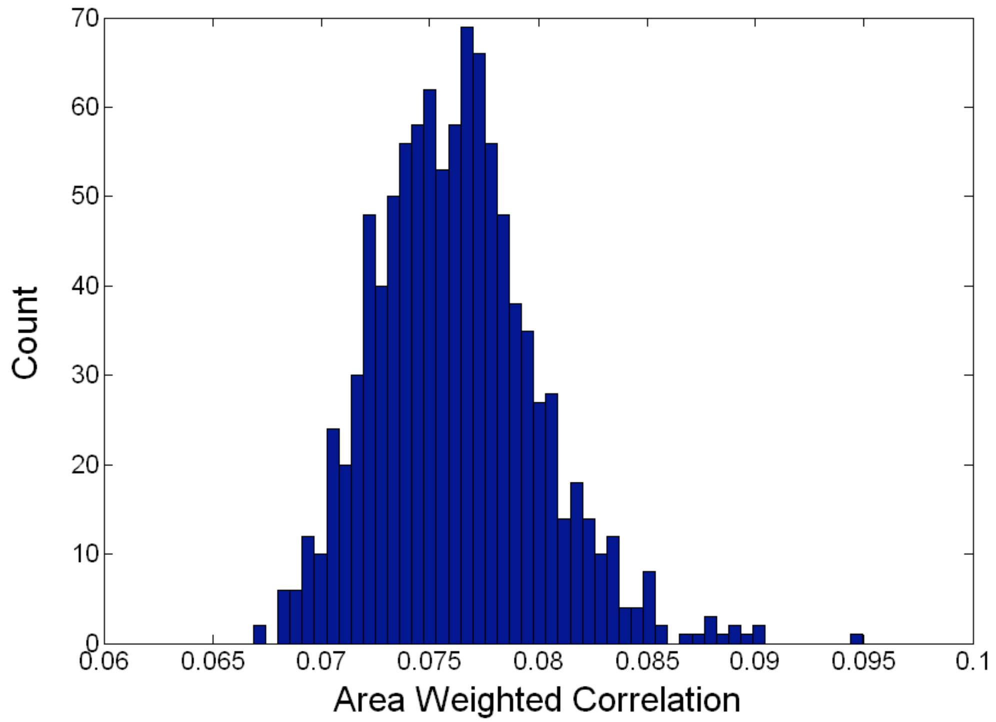


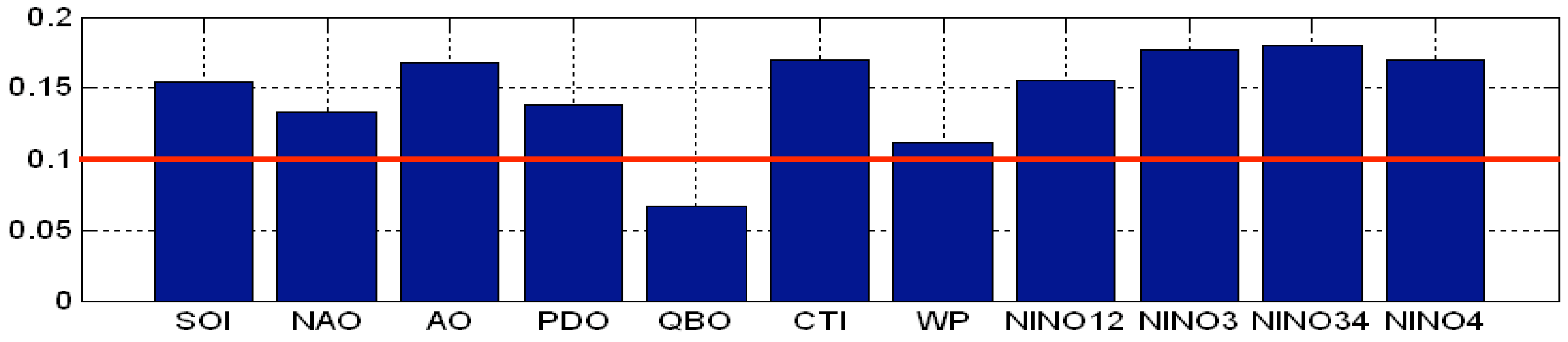
Influence of Climate Indices on Land: Area Weighted Correlation

- Correlation of an index with a land variable is a standard way to evaluate its “influence.”
  - Correlation does not imply causality.
  - Temperature and precipitation are the typical land variables.
- If relatively many land points have a relatively high correlation, then an index is influential.
- To evaluate whether clusters (or pairs) are potential indices we compute their area weighted correlation.
  - Weighted average of the correlation with land points, where weight is based on area.
  - May exclude points whose correlation is low and then calculate area weighted correlation.

Baseline for Area Weighted Correlation

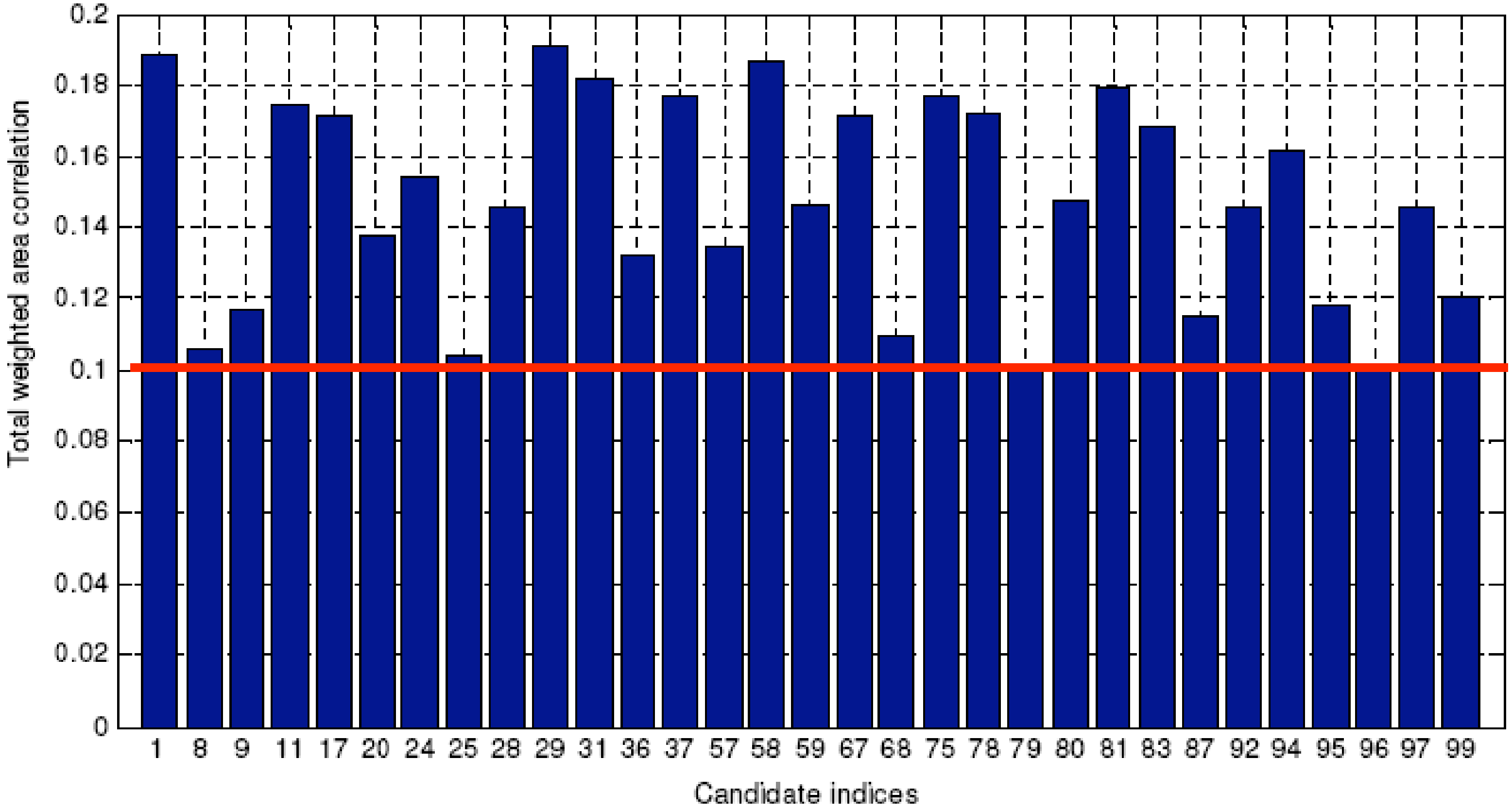
- Need to establish what level of area weighted correlation is significant
  - Baseline based on correlation of random time series to land temperature
  - Typical values of current indices





Evaluating Cluster Centroids as Potential Climate Indices

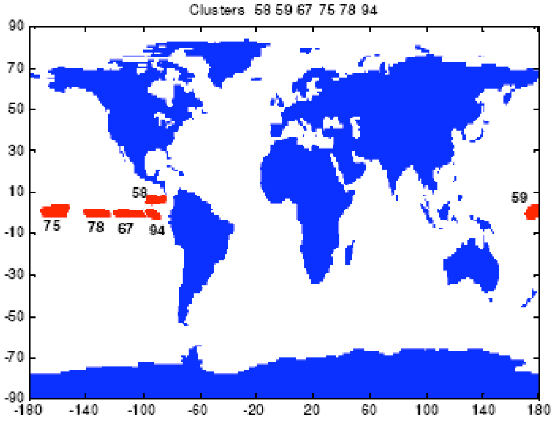
- Evaluation will be based on area weighted correlation



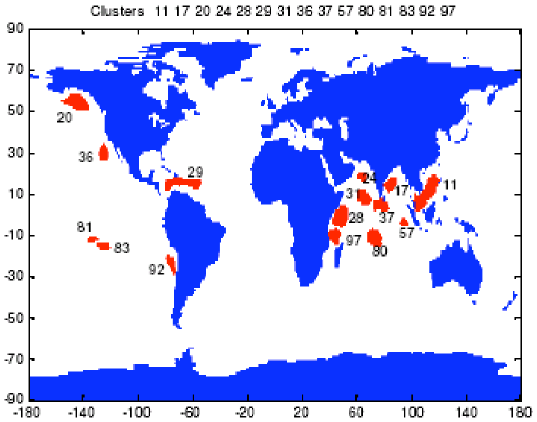
SST Clusters With Area Weighted Correlation > 0.1

Evaluating Cluster Centroids as Potential Climate Indices

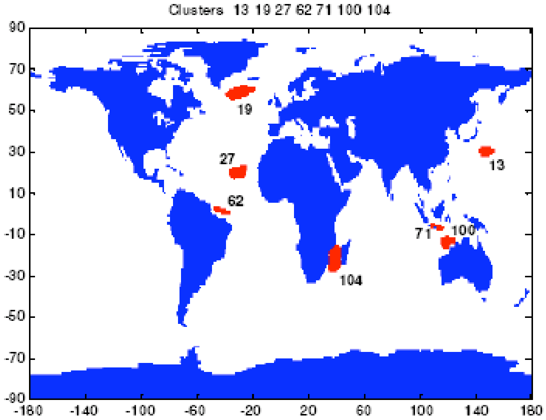
- Four cases based on similarity to known indices:



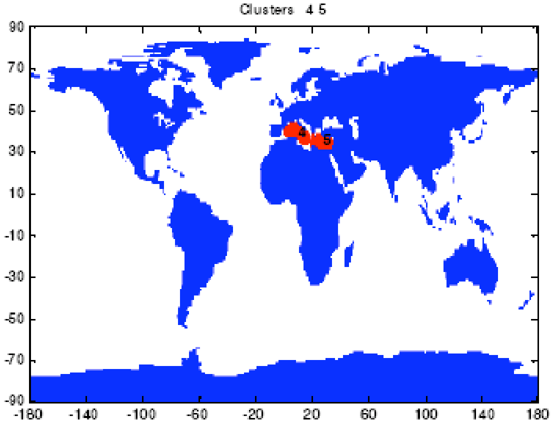
Highly similar to known indices (corr ≥ 0.8)



Similar to known indices (0.4 ≤ corr < 0.8)

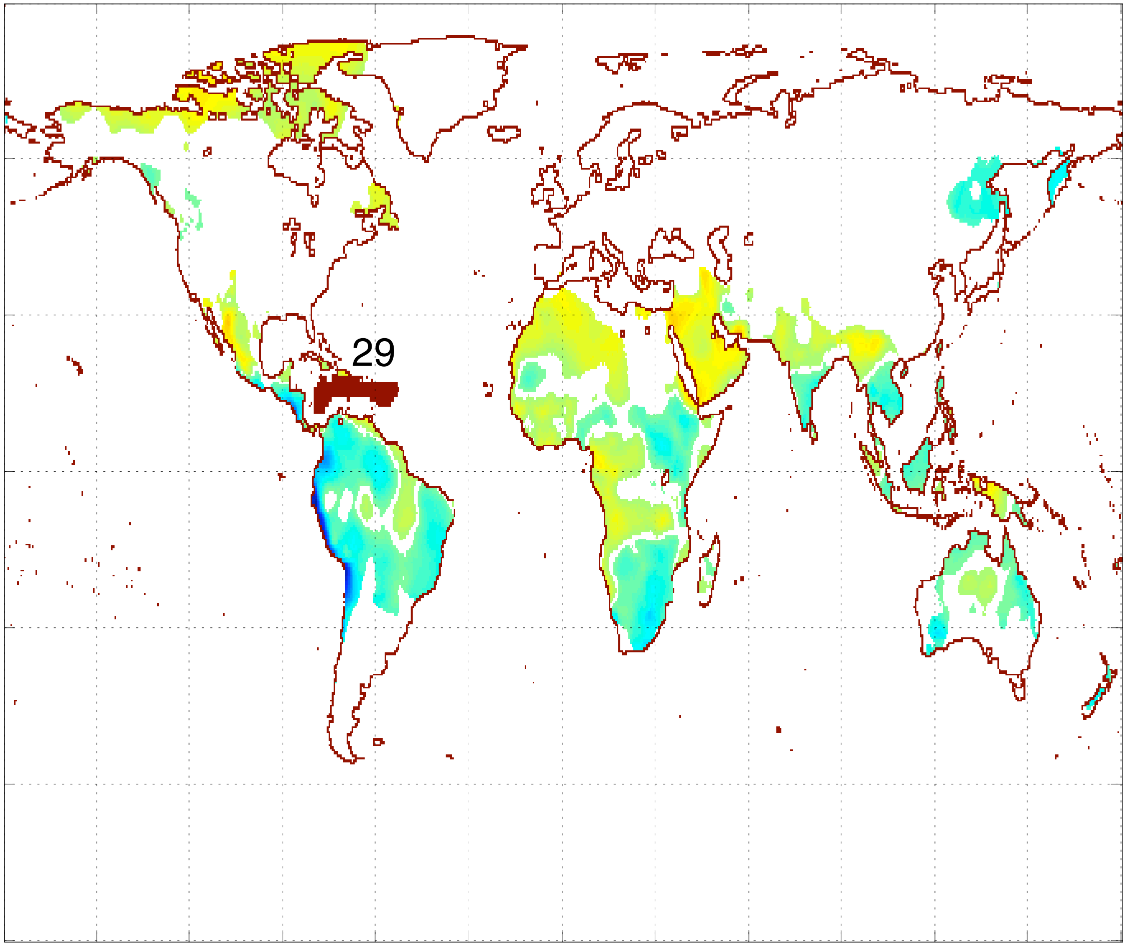


Slightly similar to known indices (0.25 ≤ corr < 0.4)

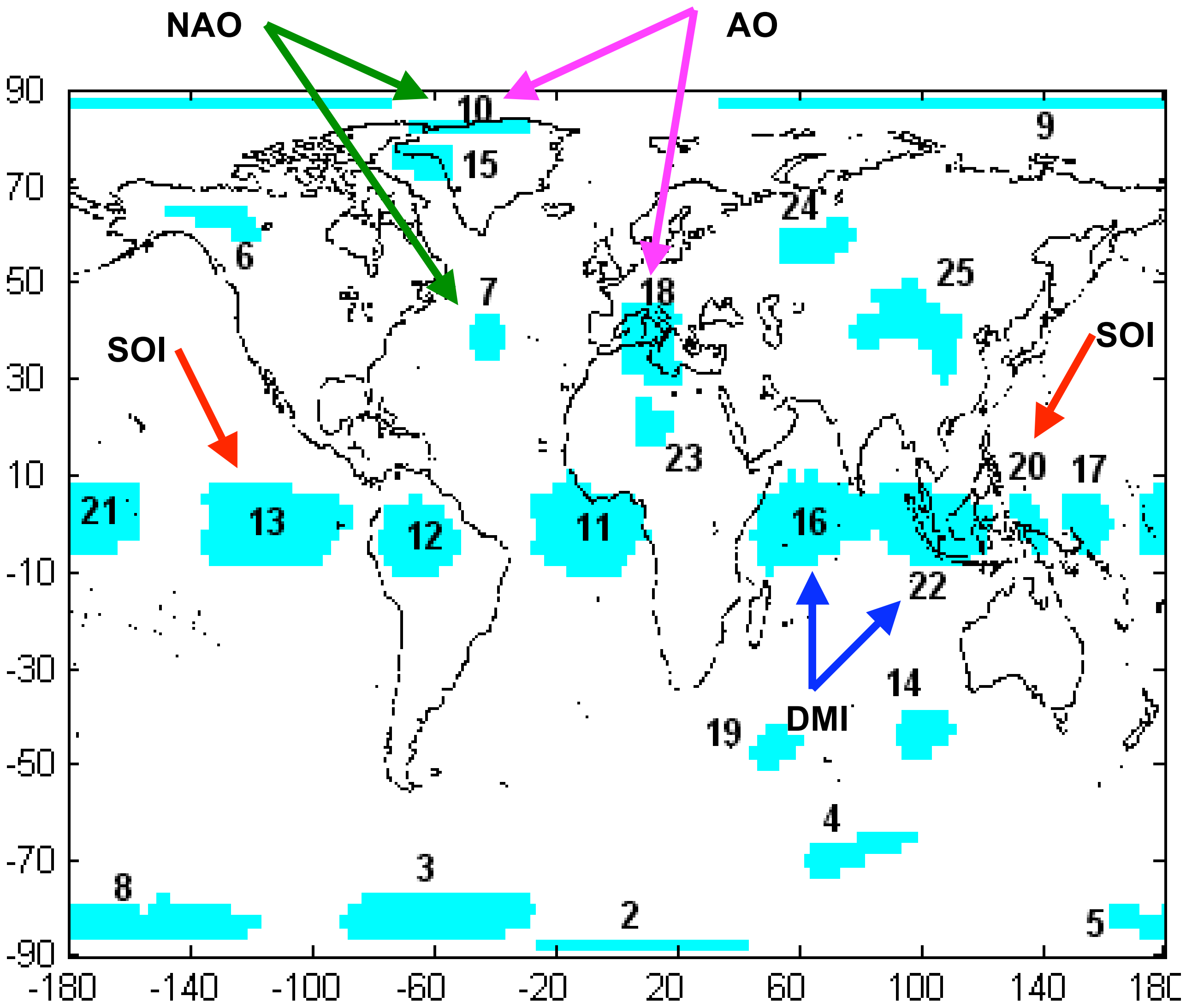


Not very similar to known indices (corr < 0.25)

An SST Cluster Moderately Correlated to Known Indices

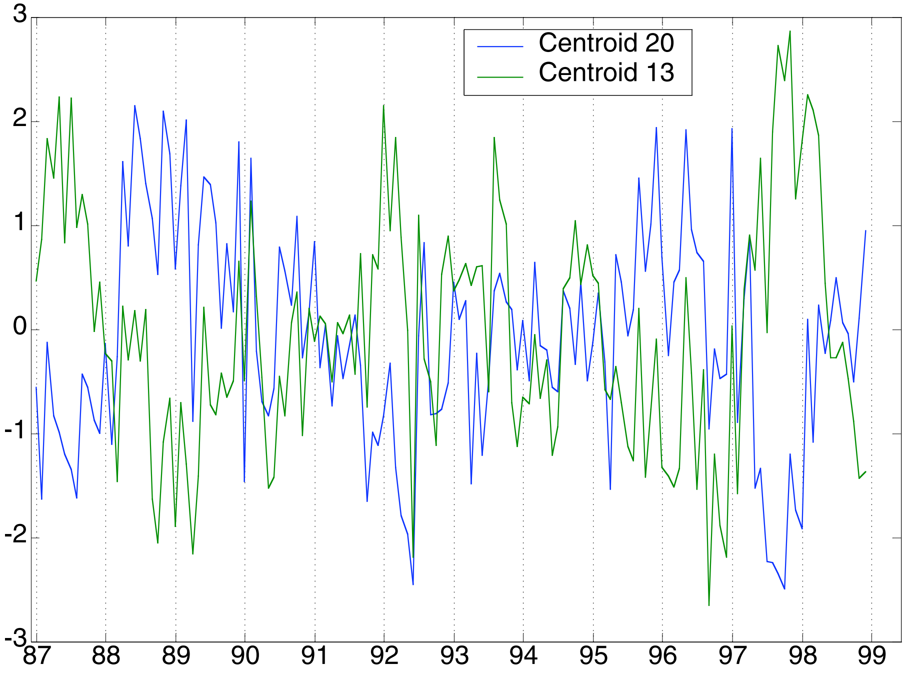


SLP Clusters

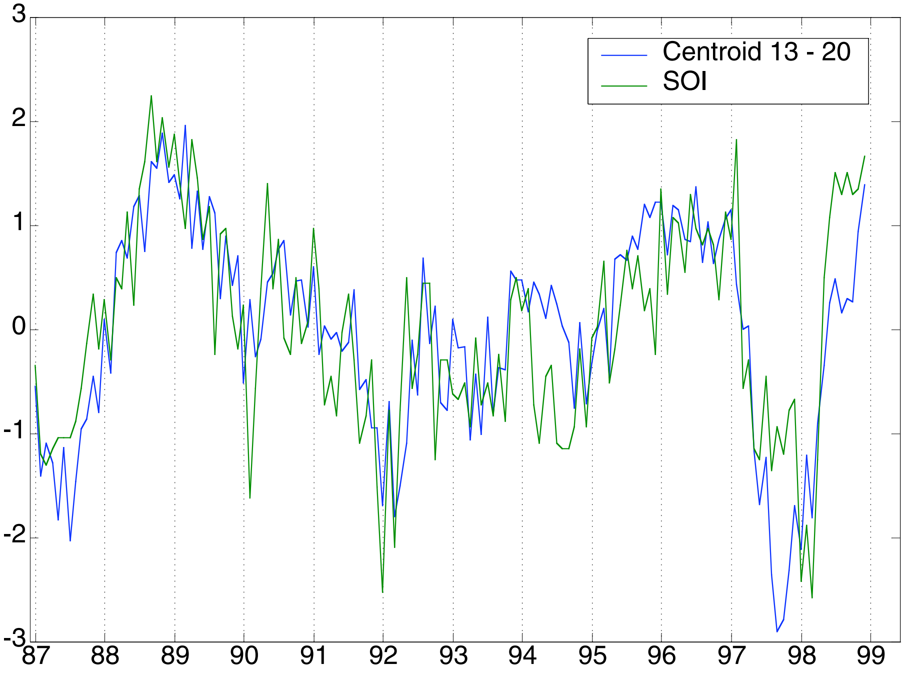


Pair of SLP Clusters that Correspond to SOI

Centroids of SLP clusters 13 and 20



Cluster centroid 20 – 13 versus SOI



Correlation = 0.75

Correlation of Known Indices with SST and SLP Cluster Centroids and SVD Components

Climate Indices	Cluster Centroids		SVD Components	
	Best-shifted Correlation	Best SST Centroid or SLP Pair	Best-shifted Correlation	Best SVD Component
SOI	-0.73	c13 - c20	0.61	3 (SLP)
NAO	0.75	c7 - c10	0.60	2 (SLP)
AO	-0.76	c10 - c18	0.82	2 (SLP)
PDO	0.52	20	-0.47	7 (SST)
QBO	-0.27	20	0.32	11 (SST)
CTI	0.91	67	-0.63	3 (SST)
WP	-0.29	c13 - c20	0.27	11 (SLP)
NINO1+2	0.92	94	-0.54	1 (SST)
NINO3	0.95	67	-0.65	1 (SST)
NINO 3.4	0.92	78	-0.68	1 (SST)
NINO 4	0.92	75	-0.69	1 (SST)

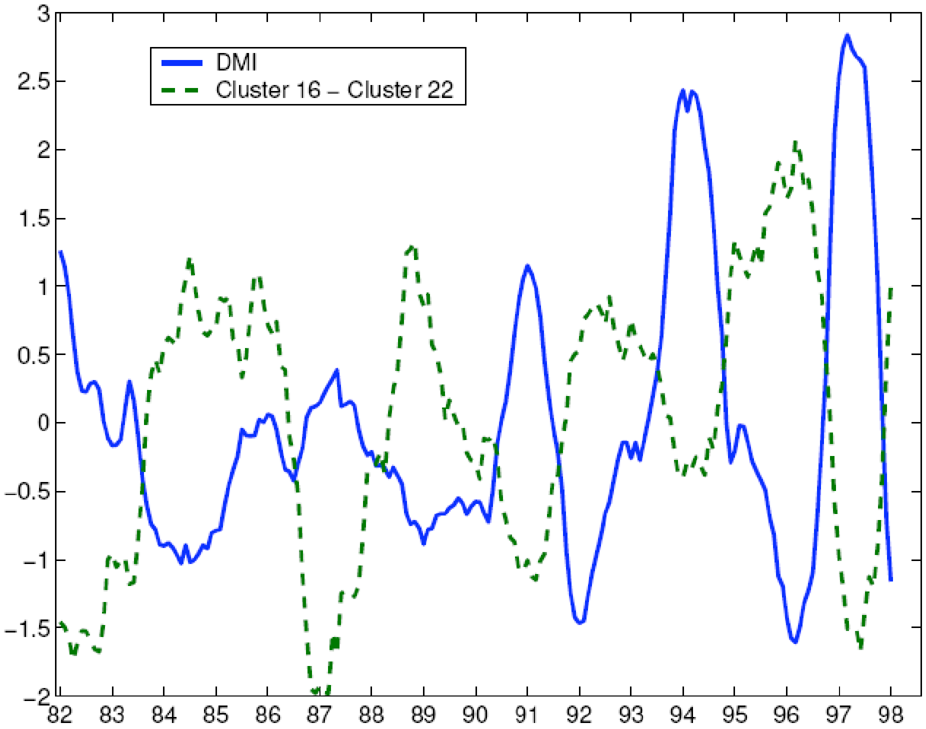
Red indicates higher magnitude of correlation.

SVD components do not have as good correlation as the cluster centroids or centroid pairs.

With some of the El Nino Indices, the leading SVD component mixes some of the indices.

Finding New Patterns: Indian Monsoon Dipole Mode Index

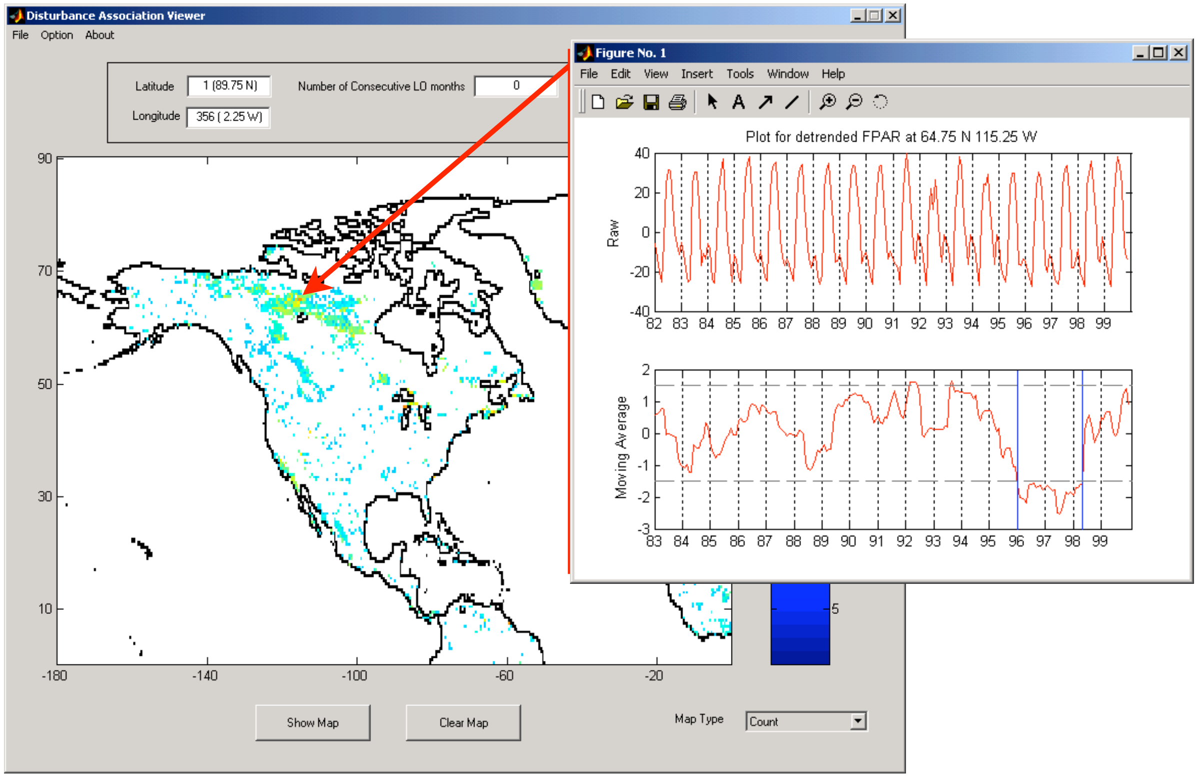
- Recently a new index, the Indian Ocean Dipole Mode index (DMI), has been discovered.
- DMI is defined as the difference in SST anomaly between the region 5S-5N, 55E-75E and the region 0-10S, 85E-95E.
- DMI and is an indicator of a weak monsoon over the Indian subcontinent and heavy rainfall over East Africa.
- We can reproduce this index as a difference of pressure indices of clusters 16 and 22.



Plot of cluster 16 – cluster 22 versus the Indian Ocean Dipole Mode index. (Indices smoothed using 12 month moving average.)

Detection of Ecosystem Disturbances

- Can detect ecosystem disturbances by detecting sudden changes in “greenness” from satellite data
  - FPAR: Fraction of Photosynthetic Active Radiation absorbed by the green part of vegetation.



Latitude: 1 (89.75 N)
Longitude: 356 (2.25 W)
Number of Consecutive LO months: 0
Map Type: Count

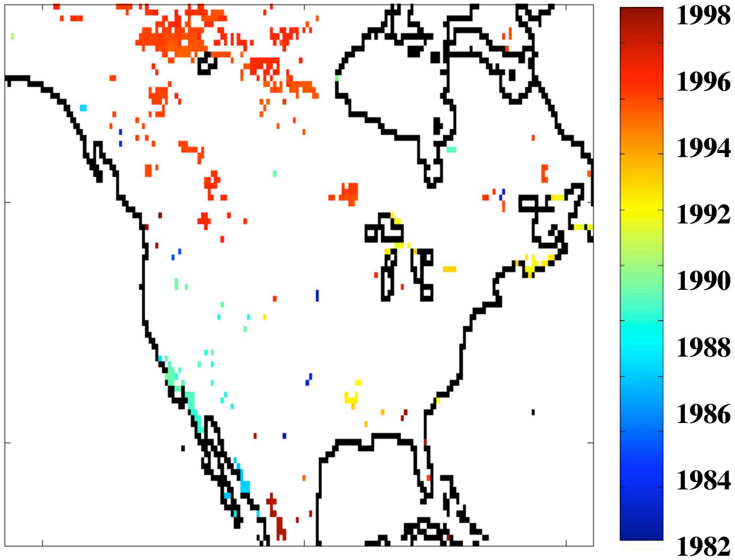
Detection of ecosystem disturbances from large global satellite data sets required development of automated techniques.

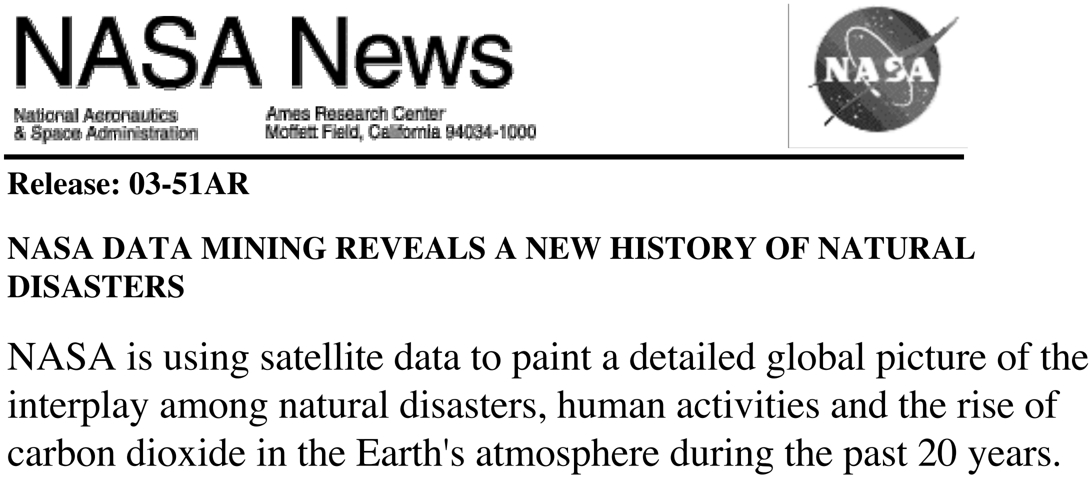
Earth Science researchers have gained deeper insight into the interplay among natural disasters, human activities and the rise of carbon dioxide in Earth's atmosphere during two recent decades.

Detection of Ecosystem Disturbances

Major ecosystem disturbances detected in North America.

NASA image of patterns in the 18-year record (1982-1999) of global satellite observations of vegetation greenness from the Advanced Very High Resolution Radiometer (AVHRR). Different colored areas identify the major ecosystem disturbance events detected and the year they occurred. The majority of potential disturbance events pictured occurred in boreal forest ecosystems of Canada or shrublands and rangelands of the southern United States.





NASA News

National Aeronautics
& Space Administration

Ames Research Center
Moffett Field, California 94034-1000

Release: 03-51AR

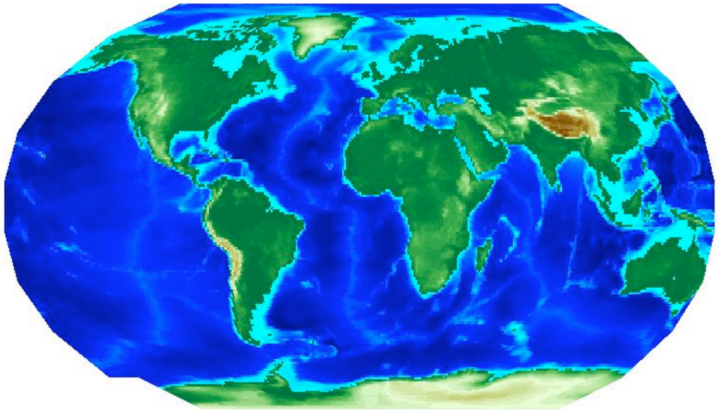
NASA DATA MINING REVEALS A NEW HISTORY OF NATURAL DISASTERS

NASA is using satellite data to paint a detailed global picture of the interplay among natural disasters, human activities and the rise of carbon dioxide in the Earth's atmosphere during the past 20 years.

Potter, C., Tan, P., Steinbach, M., Klooster, S., Kumar, V., Myneni, R., Genovese, V., 2003. Major disturbance events in terrestrial ecosystems detected using global satellite data sets. *Global Change Biology*, July, 2003.

http://amesnews.arc.nasa.gov/releases/2003/03_51AR.html

Mining Associations in Earth Science Data: Challenges



Transaction Id	*Items*
1	Bread, Milk
2	Beer, Diaper, Bread, Eggs
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Bread, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

- How to transform Earth Science data into transactions?
  - What are the “baskets”?
  - What are the “items”?
  - How to define “support”?

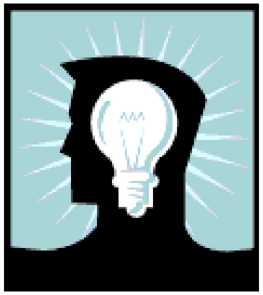
Mining Associations Patterns in Earth Science Data: Challenges

(Lat,Long,time)	*Events*
(10N,10E,1)	{Temp-Hi, Prec-Lo}
(10N,10E,2)	{Temp-Hi,Prec-Lo,NPP-Lo}
(10N,11E,2)	{Temp-Hi, NPP-Lo}
(10N,11E,5)	{Solar-Hi, NPP-Lo}
(10N,11E,10)	{Prec-Hi, PET-LO}

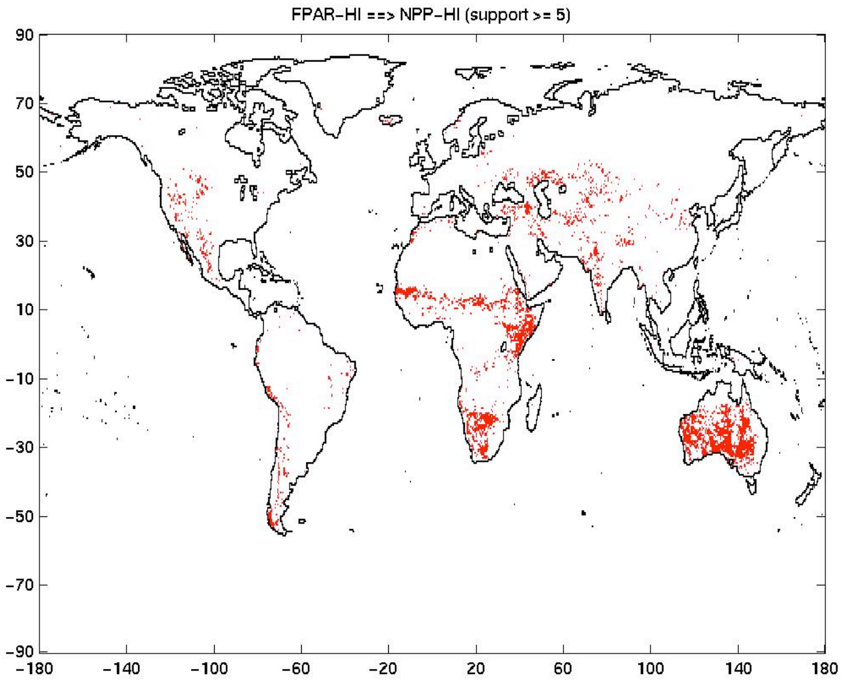
```
1 FPAR-HI PET-HI PREC-HI SOLAR-HI TEMP-HI ==> NPP-HI
            (support count=145, confidence=100%)
2 FPAR-HI PET-HI PREC-HI TEMP-HI ==> NPP-HI
            (support count=933, confidence=99.3%)
3 FPAR-HI PET-HI PREC-HI ==> NPP-HI
            (support count=1655, confidence=98.8%)
4 FPAR-HI PET-HI PREC-HI SOLAR-HI ==> NPP-HI
            (support count=268, confidence=98.2%)
                        …
```

- How to efficiently discover spatio-temporal associations?
  - Use existing algorithms.
  - Develop new algorithms.

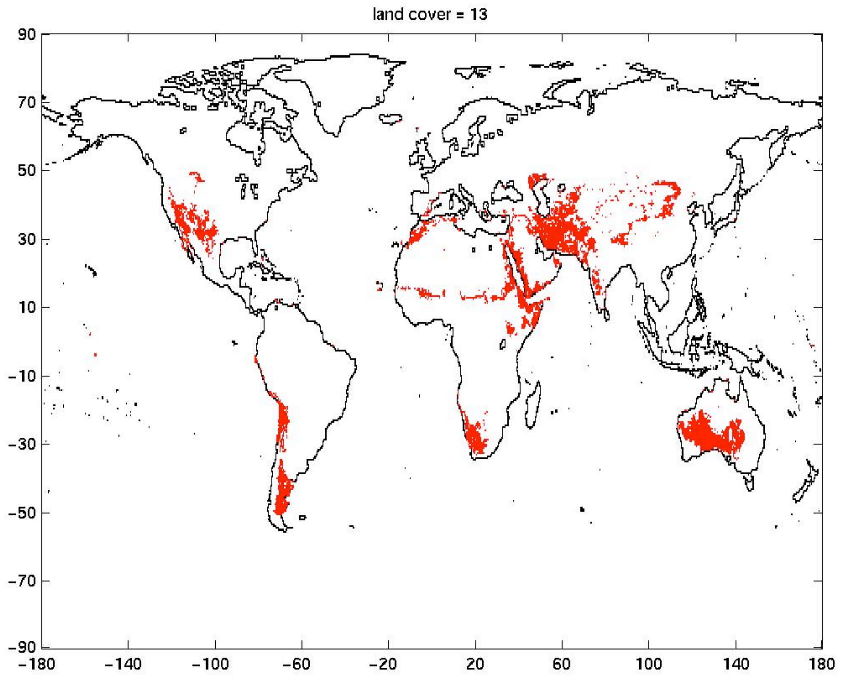
- How to identify interesting patterns?
  - Use objective interest measures.
  - Use domain knowledge.



Example of Interesting Association Patterns



FPAR-Hi ==> NPP-Hi
(sup=5.9%, conf=55.7%)



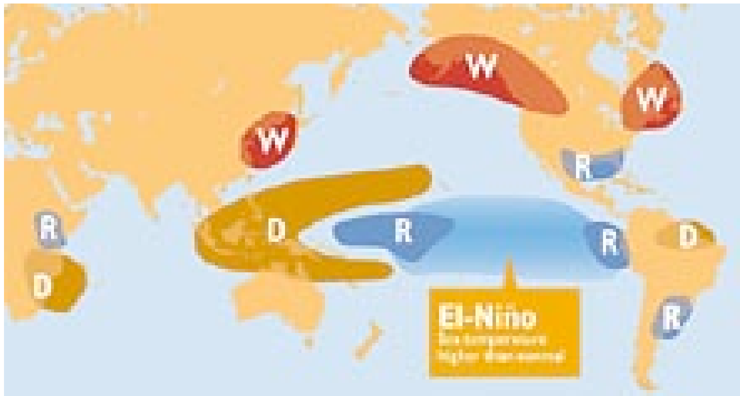
Shrubland areas

Rule has high support in shrubland areas

Efficient Query Processing Techniques in Spatial Time Series Data

Motivation

- Spatial Time Series Data
  - A collection of time series, each referring to a spatial location
  - e.g., sea surface temperature data in Pacific in 1950-2000
  - Correlation-based Similarity Queries: range query, nearest-neighbor query, join query
    - E.g., finding effected land regions in US by El Nino Index(a time series)
- Queries are Computationally Expensive!
  - Large spatial location and long time series (ts)
  - E.g., data with 1M and 2M locations with ts=600, # correlation = 1M * 2M = 2T times
- Applications
  - Data Mining in NASA Earth Science Data
    - Tele-connection Finding
    - e,.g., El Nino effects
      - Anomalous warming in Pacific
      - Heavy rainfall in Peru
      - Warming in Minnesota



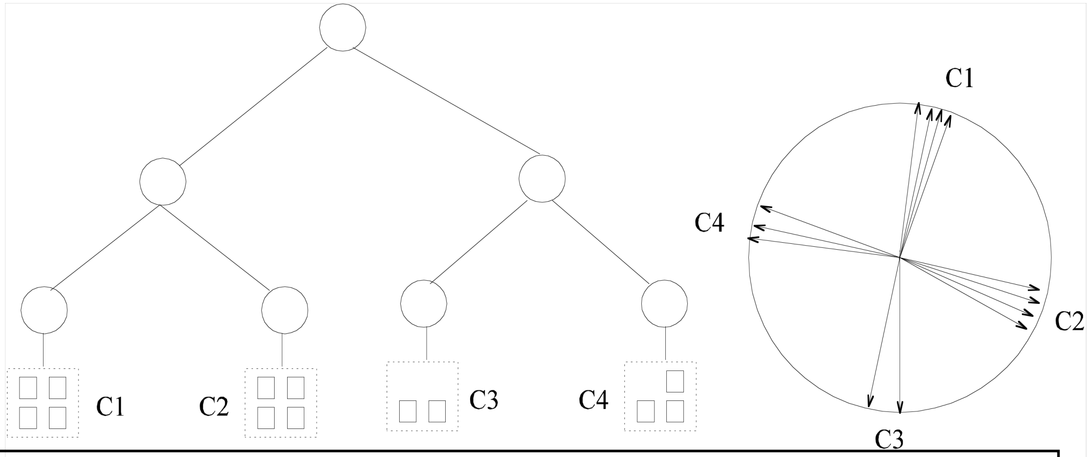
Efficient Query Processing Techniques in Spatial Time Series Data

Problem Statement

- Given:
  - A spatial time series and
  - A set of operations: e.g., range query, join, insert, delete, bulk-load
- Find:
  - A disk-based data structure
- Objective:
  - Efficiency: minimizing computational costs
- Constrains:
  - Completeness: no false dismissals for operations
  - Correctness: no false admissions for operations

Proposed Approach

- Spatial Cone tree
  - Normalized time series is located on the surface of hypersphere
  - Cone: containing multiple normalized time series in hypersphere
  - Grouping similar time series together based on spatial proximity
  - Query processing on cone-level



Efficient Query Processing Techniques in Spatial Time Series Data

Performance Evaluation

- Workload
  - NASA Earth science data
    - Monthly USA Net Primary Product data at 0.5 degree by 0.5 degree resolution in 1982-93
    - Monthly Eastern Pacific Sea Surface Temp data at 0.5 degree at 0.5 degree resolution in 1982-93
- Experimental results
  - Range Queries:
    - save 46%-89%
  - Join Queries:
    - save 40%-98%

References

[1] Pusheng Zhang, Yan Huang, Shashi Shekhar, "*Exploiting Spatial Autocorrelation to Efficiently Process Correlation-Based Similarity Queries*", Int'l Symposium on Spatial and Temporal Databases, 2003

[2] Pusheng Zhang, Yan Huang, Shashi Shekhar, "*Correlation Analysis of Spatial Time Series Datasets: A Filter-and-Refine Approach*", Pacific-Asia Conf. on Data Mining and Knowledge Discovery, 2003

[3] Pusheng Zhang, Shashi Shekhar, Vipin Kumar, Yan Huang, "*Spatial Cone Tree: An Index Structure for Correlation-based Similarity Queries on Spatial Time Series*", Int'l Workshop on Next Generation Geospatial Information

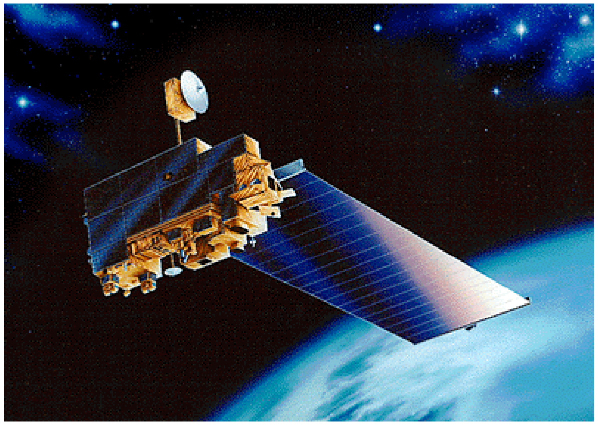
Challenges

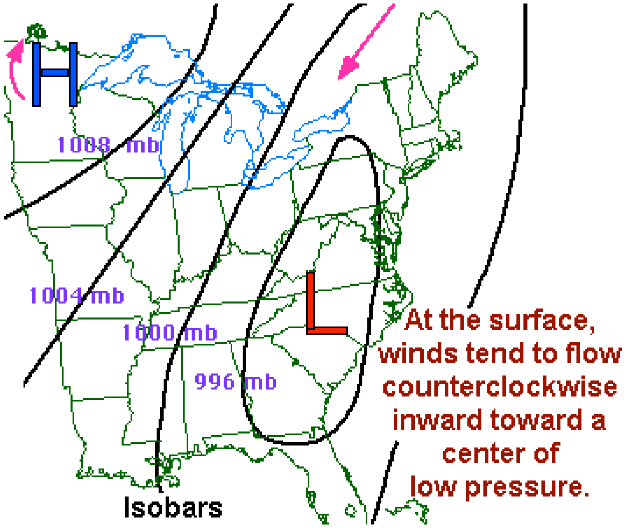
Traditional data mining techniques often work with record based data, temporal data, or spatial data, but not with data that combine all of these characteristics simultaneously.

As NASA moves to higher resolution data sets, the issues of size and high dimensionality become increasingly important.

Many ecosystem disturbances, such as forest fires, droughts, earthquakes, and volcanic eruptions, are rare events.

Relationship mining requires the selection of the "right" time periods and locations.





At the surface, winds tend to flow counterclockwise inward toward a center of low pressure.

Conclusions

- Key Innovations
  - Clustering for the detection of climate indices
  - Association analysis to discover relationships between climate variables
  - Automated detection of ecosystem disturbances
  - Efficient query processing for spatial time series

- Accomplishments to date:
  - Nine papers at computer science conferences, six peer-reviewed Earth Science journal papers, one book chapter, and a recent NASA Ames press release.

Discovery of Changes from the Global Carbon Cycle and Climate System Using Data Mining:

Journals

Disturbance Analysis

Christopher Potter, Pang-Ning Tan, Michael Steinbach, Steven Klooster, Vipin Kumar, Ranga Myneni, Vanessa Genovese, "*Major Disturbance Events in Terrestrial Ecosystems Detected using Global Satellite Data Sets*", *Global Change Biology*, 2003.

Teleconnections

C. Potter, S. Klooster, M. Steinbach, P. Tan, V. Kumar, S. Shekhar, R. Nemani, and R. Myneni, "*Global Teleconnections of Ocean Climate to Terrestrial Carbon Flux*", *J. of Geophysical Research, Vol. 108, No. D17, 4556, 2003.*

C. Potter, S. Klooster, M. Steinbach, P. Tan, V. Kumar, S. Shekhar, and C. Carvalho, "*Understanding Global Teleconnections of Climate to Regional Model Estimates of Amazon Ecosystem Carbon Fluxes*", Global Change Biology, 2003

Terrestrial Carbon Sinks

Christopher Potter , Steven Klooster, Ranga Myneni, Vanessa Genovese, Pang-Ning Tan, Vipin Kumar, *"Continental Scale Comparisons of Terrestrial Carbon Sinks", Global and Planetary Change, 39, 201-213, 2003*

C. Potter, S. Klooster, M. Steinbach, P. Tan, V. Kumar, R. Myneni, V. Genovese, "*Variability in Terrestrial Carbon Sinks Over Two Decades: Part 1-North America*", *Earth Interactions*,2003

River Analysis

Christopher Potter, Pusheng Zhang, Steven Klooster, Vanessa Genovese, Shashi Shekhar, Vipin Kumar "*Understanding the Controls of Historical River Discharge Data on Largest River Basins* ", *Earth Interactions, 2003*

Conferences

Clustering Analysis

Michael Steinbach, Pang-Ning Tan, Vipin Kumar, Christopher Potter, and Steven Klooster, "*Discovery of Climate Indices using Clustering*", KDD 2003

Michael Steinbach, Pang-Ning Tan, Vipin Kumar, Christopher Potter, and Steven Klooster, "*Temporal Data Mining for the Discovery and Analysis of Ocean Climate Indices*", *KDD Workshop on Temporal Data Mining, 2002*

Michael Steinbach, Pang-Ning Tan, Vipin Kumar, Christopher Potter, and Steven Klooster, "*Data Mining for the Discovery of Ocean Climate Indices*", *The Fifth Workshop on Scientific Data Mining (2nd SIAM International Conference on Data Mining), 2002*

Vipin Kumar, Michael Steinbach, Pang-Ning Tan, Steven Klooster, Christopher Potter, Alicia Torregrosa, "*Mining Scientific Data: Discovery of Patterns in the Global Climate System* ", *Joint Statistical Meeting, 2001*

Michael Steinbach, Pang-Ning Tan, Vipin Kumar, Christopher Potter, Steven Klooster, Alicia Torregrosa, "*Clustering Earth Science Data: Goals, Issues and Results*", *KDD Workshop on Mining Scientific Datasets, 2001*

Association Analysis

Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Christopher Potter, Steven Klooster, Alicia Torregrosa, "*Finding Spatio-Temporal Patterns in Earth Science Data*", *KDD Workshop on Temporal Data Mining, 2001*

Query Processing

Pusheng Zhang, Yan Huang, Shashi Shekhar, and Vipin Kumar, *"Exploiting Spatial Autocorrelation to Efficiently Process Correlation-Based Similarity Queries"*, the 8th *Symp. on Spatial and Temporal Databases*, 2003

Pusheng Zhang, Yan Huang, Shashi Shekhar, and Vipin Kumar, *"Correlation Analysis of Spatial Time Series Datasets: A Filter-and-Refine Approach"*, the *Seventh Pacific-Asia Knowledge Discovery and Data Mining*, 2003

Book Chapter

Pusheng Zhang, Michael Steinbach, Vipin Kumar, Shashi Shekhar, Pang-Ning Tan, Steve Klooster, and Chris Potter, Discovery of Patterns of Earth Science Data Using Data Mining, as a Chapter in *Next Generation of Data Mining Applications*, Jozef Zurada and Medo Kantardzic(eds), IEEE Press, 2003